# SWAN
# Service for
# Web-based ANalysis

*https://swan.cern.ch*

E. Tejedor, D. Piparo, P. Mató – EP-SFT
L. Mascetti, J. Moscicki, M. Lamanna – IT-ST

IML meeting

*25/08/2016*

# Prelude: The Notebook

**Notebook**: A web-based interactive computing interface and platform that combines code, equations, text and visualisations.

Many supported languages: Python, Haskell, Julia, R … One generally speaks about a "kernel" for a specific language

In a nutshell: an "interactive shell opened within the browser"

Also called:
"Jupyter Notebook" or "IPython Notebook"

# SWAN: Data analysis "as a service"

*Interface:* Jupyter Notebooks

*Goals:*

- Analysis only with a web browser
  - Platform independent ROOT-based data analysis
  - Calculations, input and results "in the Cloud"
- Easy sharing of scientific results: plots, data, code
  - Storage is crucial: mass & synchronised
- Simplify teaching of data processing and programming
  - Gallery of analysis examples
- Integration with other analysis ecosystems: R, Python, …

# ROOT-Jupyter Integration

ROOT has been fully integrated with the Jupyter technology

- ROOT C++ kernel
- Python kernel: activation via “import ROOT”
- JavaScript interactive visualisation
- Other goodies: tab completion, magics, …
- Ongoing: full TMVA integration (Google Summer of Code)
  - Enhanced JSROOT plots, interactive training, neural network visualisation, …
  - To be presented at the next IML meeting

# SWAN in the CERN Ecosystem

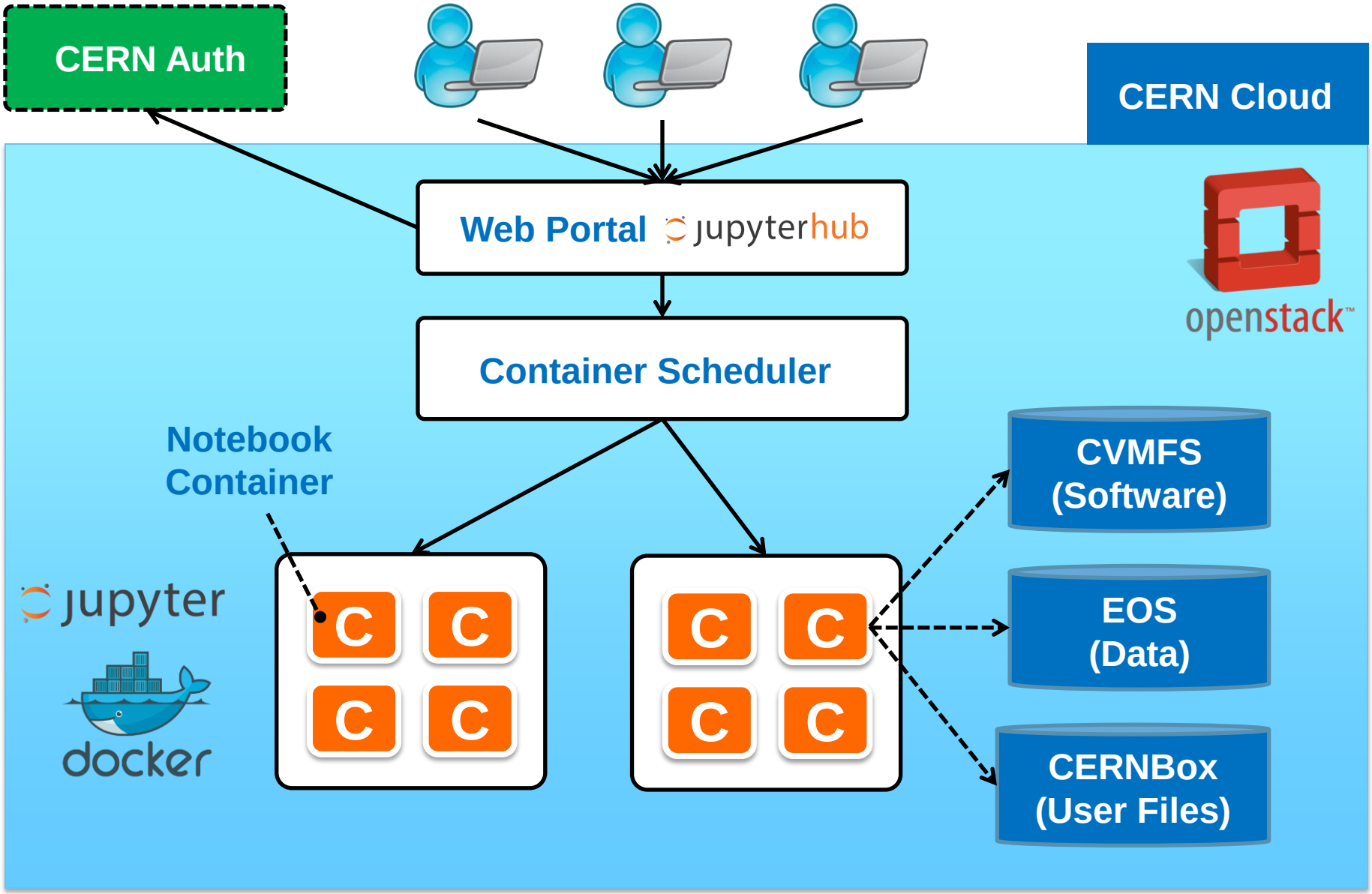
SWAN relies on production technologies at CERN:

- Authentication with CERN credentials (SSO)
- Infrastructure: virtual machines in OpenStack Cloud
- Software distribution: CVMFS
  - Centrally distributed software, managed by EP-SFT
- Storage access: CERNBox, EOS
  - All experiment data potentially available!

Plus some external technologies:

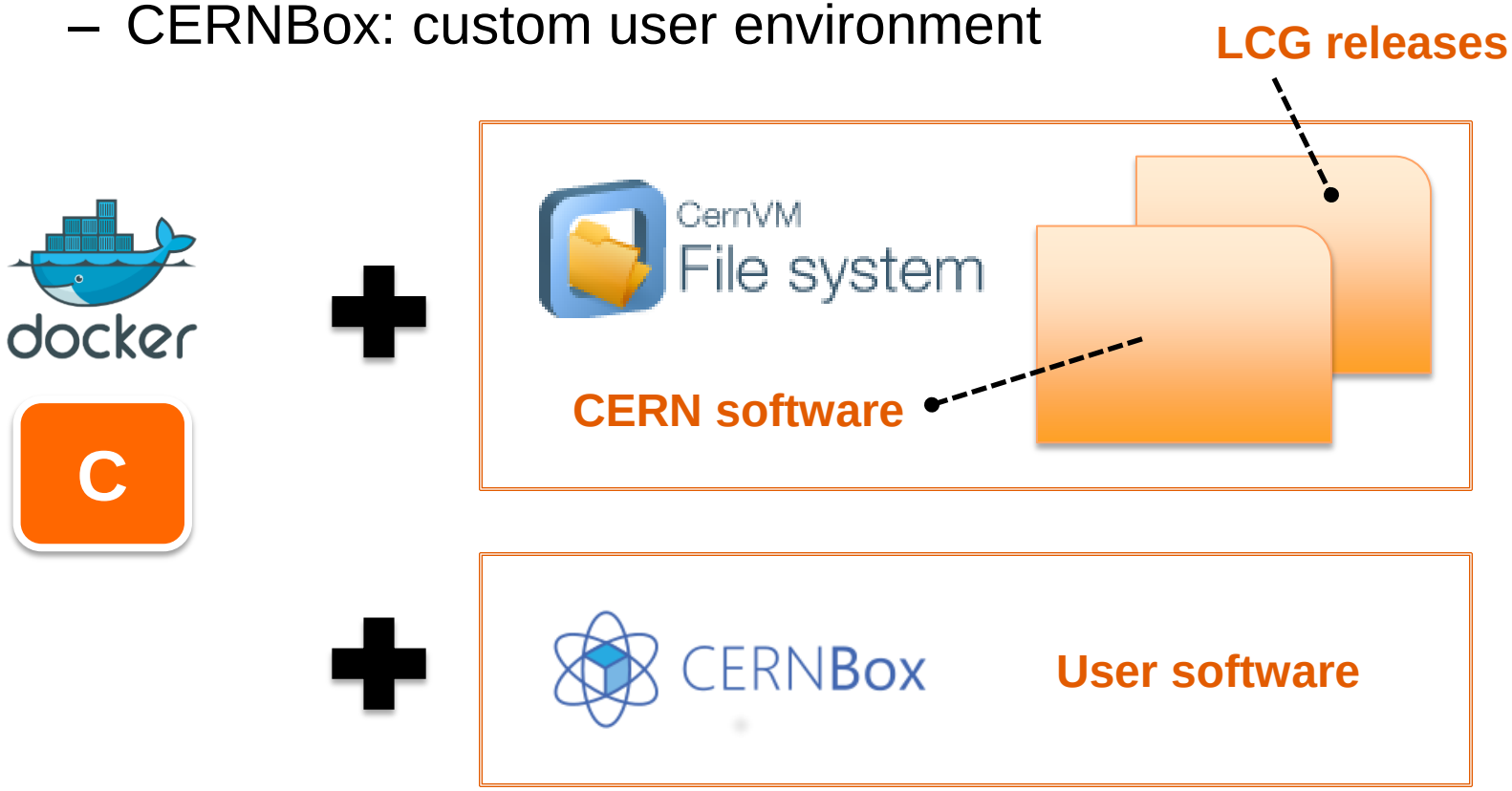
- JupyterHub
- Docker

# Service Architecture



# Software Environment

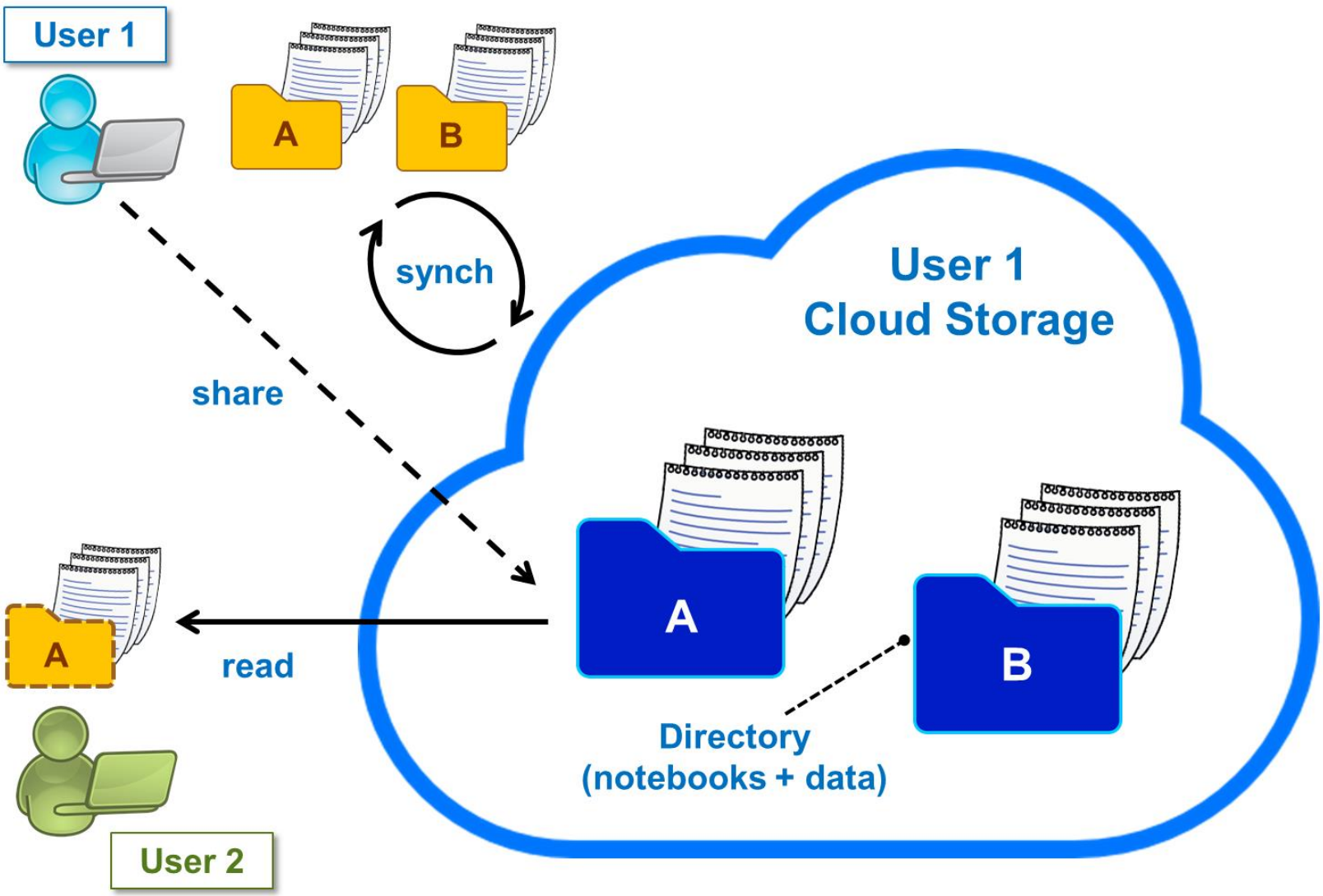
- Strategy to configure the software environment:
  - Docker: single thin image, not managed by the user!
  - CVMFS: configurable environment via “views”
  - CERNBox: custom user environment



# Software Environment: ML

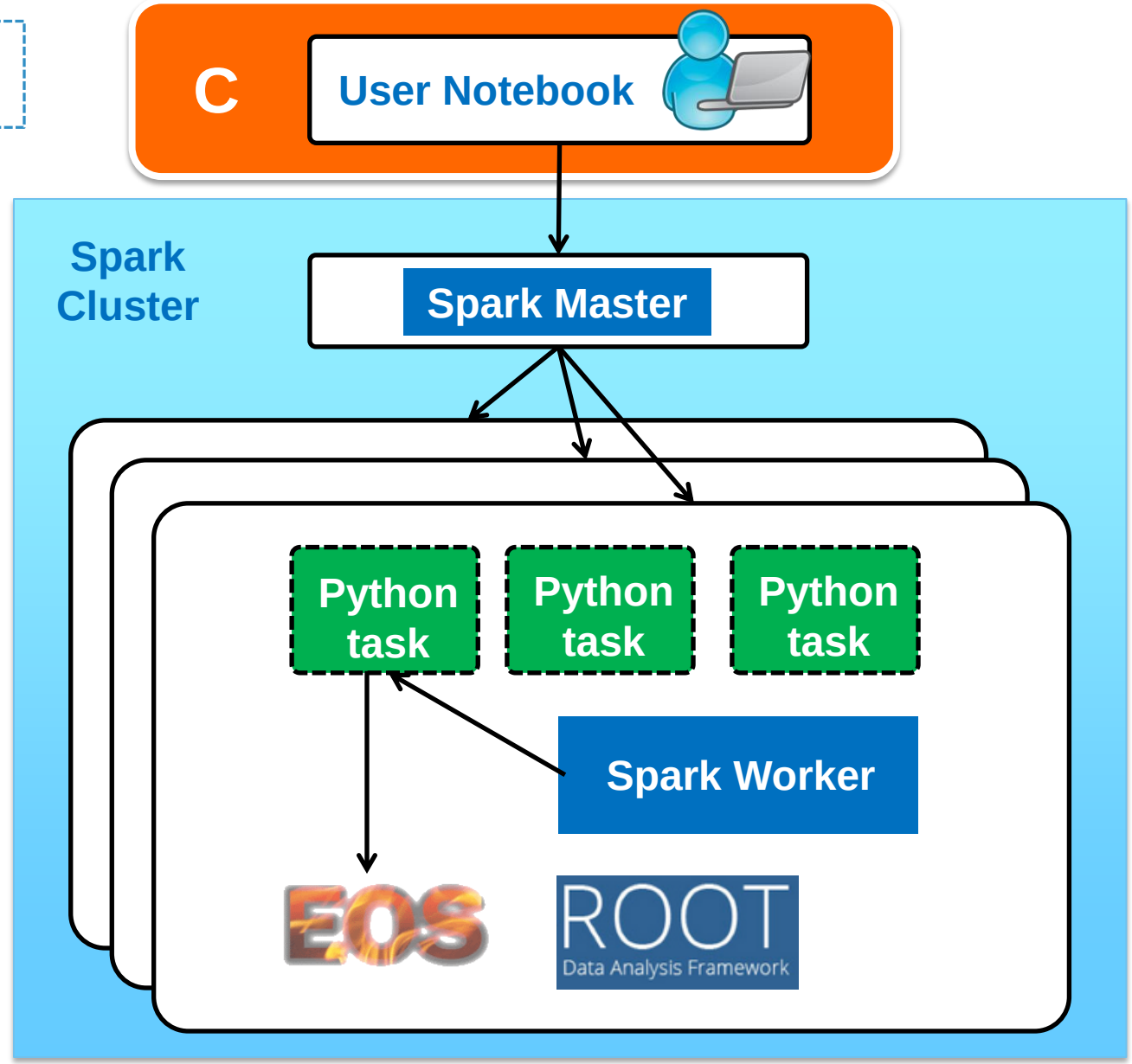
- TMVA: up to date with new features
  - e.g. CrossValidation, cleaner output, …
- Not restricted to TMVA, though
  - Scikit-learn already there
  - Spark MLlib: tutorial by IT-DB next week
- Open to incorporate new ML libraries to CVMFS
  - To be coordinated with IML and EP-SFT librarian
  - Number of users and potential impact should be enough to justify its addition
  - For testing, installation in CERNBox is possible from within SWAN
    - e.g. for Python: *pip install –user mypackage*

# CERNBox: Sync & Share



# R&D: Offloading from SWAN

In collaboration with IT-DB-SAS



# R&D: Sharing from SWAN

CMSDimuon_py Last Checkpoint: an hour ago (autosaved)

Control Panel Logout

File Edit View Insert Cell Kernel Help

Share

Python 2

Code CellToolbar

## Dimuon spectrum

This ROOTbook produces a plot of the dimuon spectrum starting from a subset of the CMS collision events of Run2010B.

Dataset Reference:
McCauley, T. (2014). Dimuon event information derived from the Run2010B public Mu dataset. CERN Open Data Portal. DOI: 10.7483/OPENDATA.CMS.CB8H.MFFA.

In [ ]:
```
import ROOT
```

A little extra: JavaScript visualisation. This command will become a magic very soon.

In [ ]:
```
%jsroot on
```

## Convert to ROOT format and analyse

First of all we convert the csv file into ROOT format, i.e. filling up a TTree data structure. But first of all we uncompress it if it's not.

In [ ]:
```
inputFileName = '../data/MuRun2010B.csv'
import os
if not os.path.exists(inputFileName):
    import gzip
    import shutil
    with gzip.open(inputFileName+'.gz', 'rb') as f_in, open(inputFileName, 'wb') as f_out:
        shutil.copyfileobj(f_in, f_out)
```

# SWAN Now

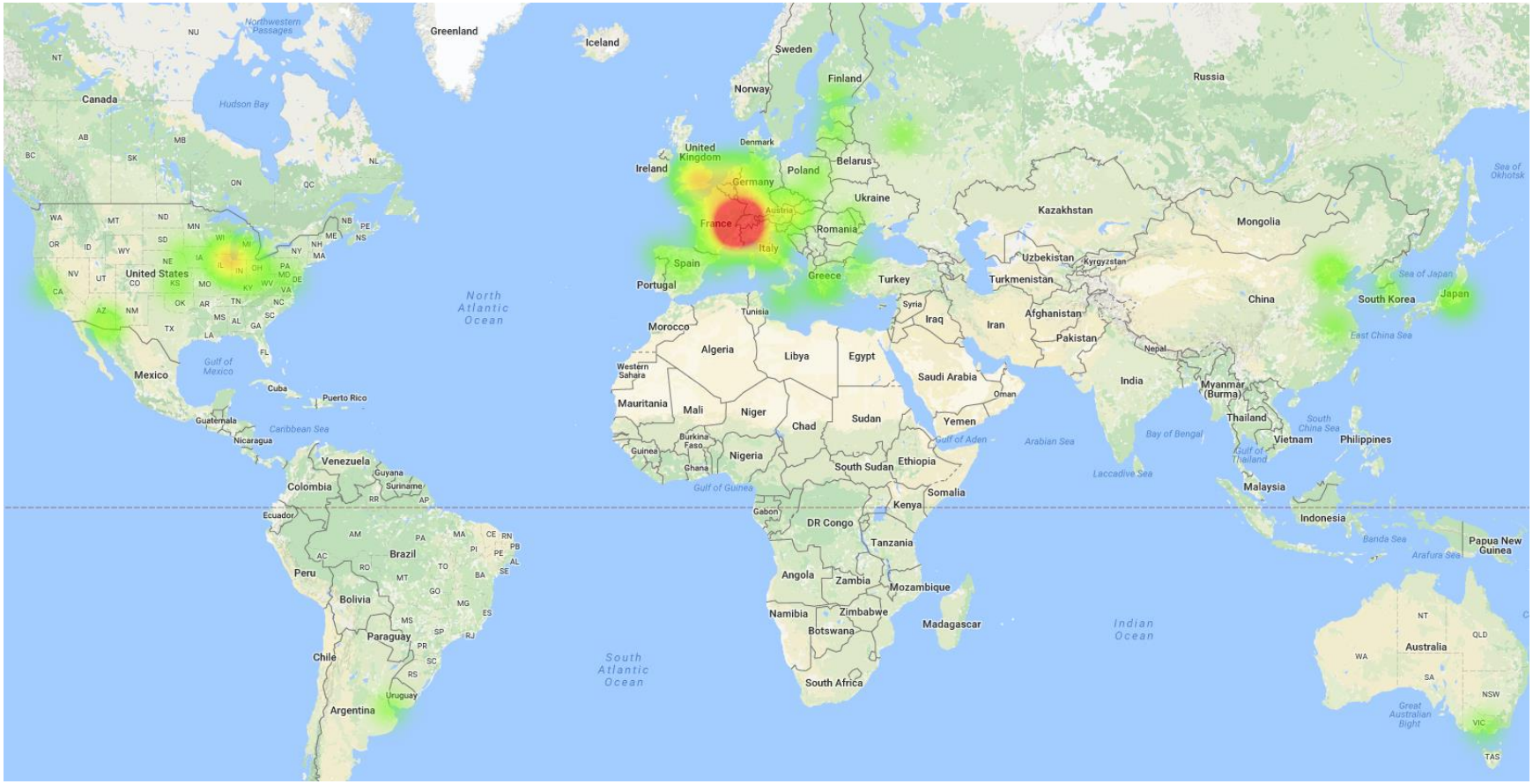
- Pilot Service released beginning of June

https://swan.cern.ch

- All the main components are already there
  - CERN SSO
  - EOS, CERNBox
  - CVMFS
- In beta testing phase: ~200 users, growing
  - If interested, please send us an e-mail to: swan-talk@cern.ch
  - Your feedback is very much welcome!!

# SWAN Now (II)

- Since a month, accessible also from outside CERN



# A Notebook Gallery

**SWAN**

Interactive Data Analysis, in the Cloud.

Set of example notebooks in **swan.web.cern.ch**

Home | Galleries | FAQ | Talks and Publications

Basic | ROOT Primer | Accelerator Complex | Machine Learning | Apache Spark

Click to open them in SWAN!

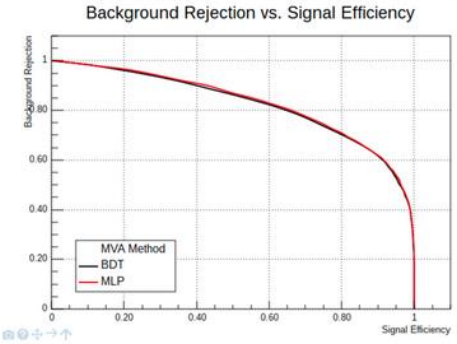
## Machine Learning

### TMVA Basics

**Plot ROC Curve**

```
In [8]: %jsroot
        TCanvas *c1=factory->GetROCCurve(loader1);
        c1->Draw();
```

Open in SWAN




### RMVA

Open in SWAN

ROOT Data Analysis Framework

R

TMVA

### Cross Validation

**Declare DataLoader**

```
In [2]: TMVA::DataLoader *loader=new TMVA::DataLoader("dataset");

        loader->AddVariable( "var1", 'F' );
        loader->AddVariable( "var2", 'F' );
        loader->AddVariable( "var3", 'F' );
        loader->AddVariable( "var4", 'F' );
        loader->AddVariable( "var5 := var1-var2", 'F' );
```

Open in SWAN

### Variable Importance

**Variable importance**

```
In [4]: TH1F* importance = factory->EvaluateImportance(loader,TMVA::VIType::kAll,TMVA::Type

        --- Variable Importance Results (All)
        --- var1 = 25.0518 %
        --- var2 = 8.55669 %
        --- var3 = 15.2109 %
        --- var4 = 32.879 %
        --- var1*var2 = 18.3016 %
```

Open in SWAN

# Summary, Plans

- Prototype service available
  - ROOT integrated with Jupyter
  - CVMFS for software distribution
  - EOS mass storage + CERNBox synchronisation
- Future plans:
  - Continue to incorporate user feedback
  - Improve experience with storage: sharing
  - Exploit external resources (e.g. Spark clusters, batch, Grid resources)
  - Provide necessary support for ML use cases
  - Open to all users

# Backup

# SWAN Now (III)

- Since a month, accessible also from outside CERN